國立政治大學111學年度第二學期 期中R程式考題

Department:____民三____ ID:__109209017_____ Name:_張育嘉_

Subject: **Regression Analysis (I)**

Date: 2023/04/20

Time: 11:00~12:00 (60 minutes)

95

注意事項:

1. 本次考題以R程式(Rgui或RStudio)方式作答，其他程式不允許。
2. 考試過程中可查詢書本、教學講義或上網，禁止利用messenger, IG, Line等等通訊軟體。
3. 禁止疑似作弊行為。
4. 本答案卷上請務必於　　內複制「執行後的程式碼及結果(含圖形)」，於本答案卷貼上(Courier New, 10點字，白底黑字)，不能只有程式碼，不能只有報表。最後，將每小題之答案(不能只印出報表，要助教去找答案)，在小題最後以打字(英文)作答(Times New Roman, 12點字，白底黑字)。
5. 請依序註明題號: (1)a, (1)b, (2)a 等等。
6. 作答完請將此word檔存檔，檔名為「學號-姓名-Regression-R-Midterm.docx」(更改成自己「學號、姓名」)並上傳至http://ftp.hmwu.idv.tw:8080/login.html?lang=tchinese
   或點選教師網站首頁【作業考試上傳區】。
7. 帳號: reg111，密碼: 上課教室號碼，資料夾: 「**20230420-MidtermExam**」
8. 如果上傳網站出現「空白頁」，請將滑鼠移至「網址列」後，按「Enter」即可。若再不行，請換其它瀏覽器(IE/Edge/Firefox/Chrome)
9. 上傳檔案無法刪除，若要上傳更新檔，請於主檔名後加

「-2」，例如:「學號-姓名-Regression-R-Midterm-2.docx」。

Notes:

1. This is an Open Book exam; you are free to use any materials including laptop, tablet and internets.
2. Smart phone and the communication software/APP (e.g., Messenger, IG, LINE, WeChat,..) are prohibited.
3. Copy the R codes and the results from and paste it to this answer sheet.
4. Change the file name of this answer sheet according to your ID and Full Name. Upload the answer sheet to http://ftp.hmwu.idv.tw:8080/login.html?lang=tchinese
5. Account: `rege111`, password: classroom number.

(1) 20% Data file: Grade_Point_Average.csv

Grade point average. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	118	119	120
X_i:	21	14	28	...	28	16	28
Y_i:	3.897	3.885	3.778	...	3.914	1.860	2.948

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.
b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?
c. Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

+20

#(1)a

```
> setwd("~/Downloads/Reg111-2_Rcode_1-3/data")
> gpa <- read_csv("~/Desktop/data/Grade_Point_Average.csv")
Rows: 120 Columns: 2

── Column specification ──────────────────────────────────────

Delimiter: ","
dbl (2): GPA, ACT

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to
```

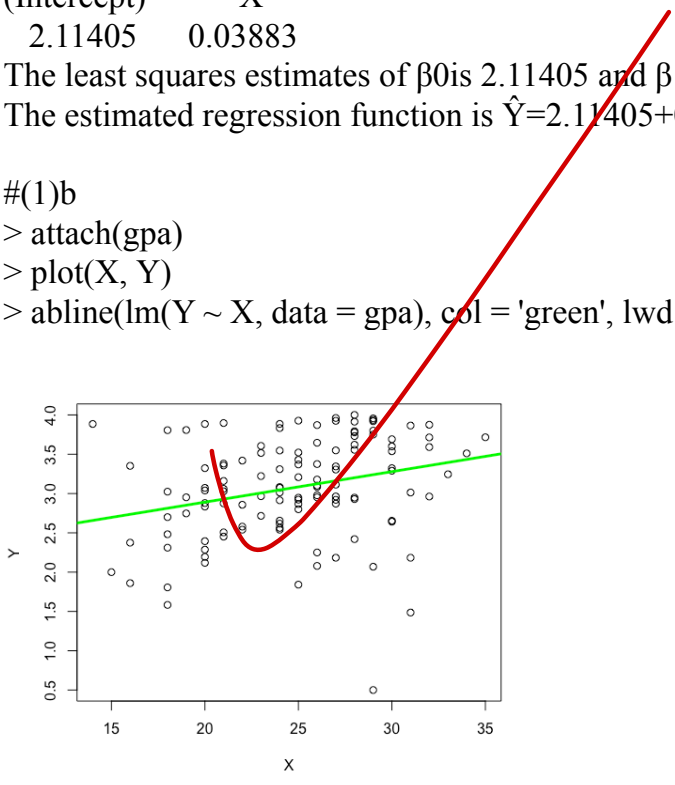
quiet this message.
> colnames(gpa) <- c("Y", "X")
> lm(Y ~ X, data = gpa)

Call:
lm(formula = Y ~ X, data = gpa)

Coefficients:
(Intercept) X
2.11405 0.03883

The least squares estimates of β0is 2.11405 and β1is 0.03883.
The estimated regression function is $\hat{Y}$=2.11405+0.03883X

#(1)b
> attach(gpa)
> plot(X, Y)
> abline(lm(Y ~ X, data = gpa), col = 'green', lwd = 3)



From the above picture, the estimated regression function appear to fit the data well.

#(1)c
> hat_y <- 2.11405 + 0.03883 * 30
> hat_y
[1] 3.27895
> gpa.lm <- lm(Y ~ X, data = gpa)
> predict(gpa.lm, data.frame(X = 30))
1
3.278863

+2

(2) Data file: Grade_Point_Average.csv
20% Refer to **Grade point average**

a. Set up the ANOVA table.
b. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.
c. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model?

#(2)a
> gpa.lm <- lm(Y ~ X, data = gpa)
> summary.aov(gpa.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	3.59	3.588	9.24	0.00292 **
Residuals	118	45.82	0.388		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#(2)b
Hypothesis Test:
H0:β1=0
H1:β1≠0
Significance level: α=0.01
Decision: From (a), the p-value of F test is 0.00292<α=0.01, reject

H0.
Conclusion: X and Y has a linear association.

#(2)c

```
> gpa.lm <- lm(Y ~ X, data = gpa)
> summary(gpa.lm)
Call:
lm(formula = Y ~ X, data = gpa)

Residuals:
    Min      1Q  Median      3Q     Max
-2.74004  -0.33827  0.04062  0.44064  1.22737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
X            0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

The absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model is 3.588 which is obtained by SSR.

(3) Data file: Grade_Point_Average_X.csv

30% Refer to **Grade point average**

a. Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?

b. Prepare a dot plot of the residuals. What information does this plot provide?

c. Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?

e. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X < 26$, $X \geq 26$, and use $\alpha = .01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

f. Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

i:	1	2	3	...	118	119	120
X_2:	122	132	119	...	140	111	110
X_3:	99	71	75	...	97	65	85

+5

#(3)a

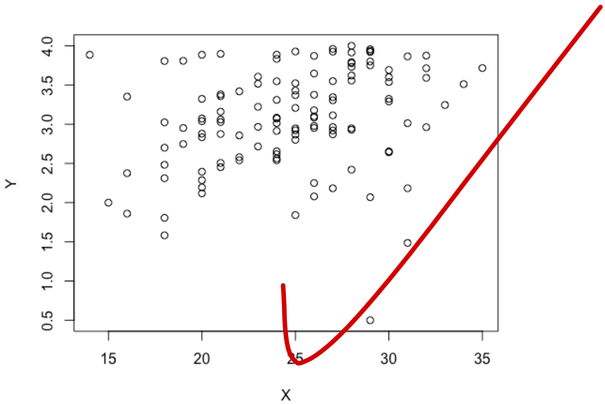
> gpa_x <-
read_csv("~/Desktop/data/Grade_Point_Average_X.csv")
Rows: 120 Columns: 4

── Column specification ──

Delimiter: ","
dbl (4): GPA, ACT, Intelligence, RankPercentile

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
> attach(gpa_x)
> plot(X, Y)
```



#3(b)

(4) Data file: Solution_concentration.csv

30% **Solution concentration.** A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The results follow.

+30

i:	1	2	3	...	13	14	15
X_i:	9	9	9	...	1	1	1
Y_i:	.07	.09	.08	...	2.84	2.57	3.10

a. Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

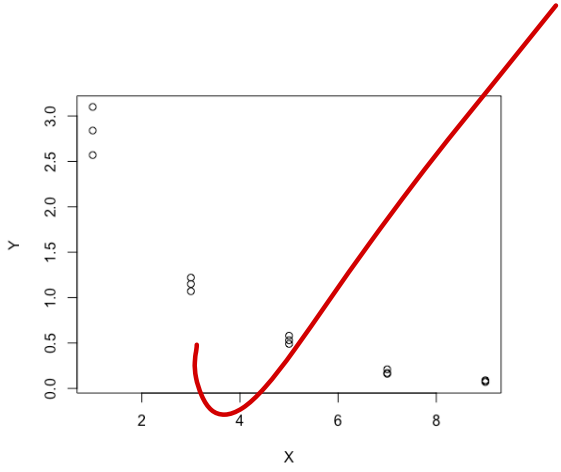
c. Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

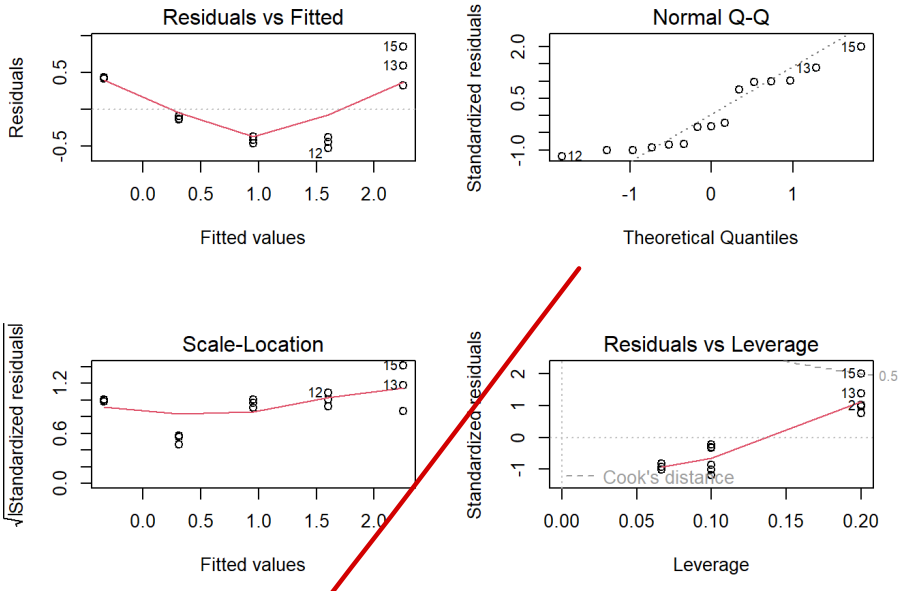
f. Express the estimated regression function in the original units.

```
#a
> concentration <-
read.csv("~/Desktop/data/Solution_concentration.csv", header = F)
> colnames(concentration) <- c("Y", "X")
> attach(concentration)
> plot(X, Y)
```



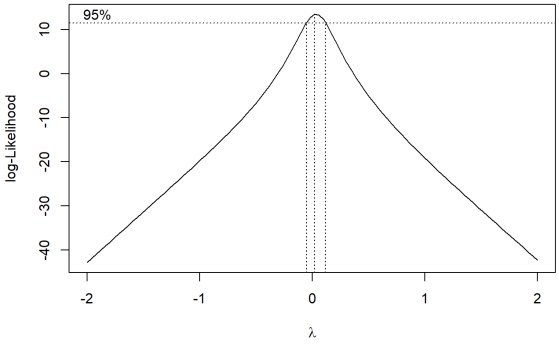

```
> concentration.lm <- lm(Y ~ X, data = concentration)
> summary(concentration.lm)
Call:
lm(formula = Y ~ X, data = concentration)

Residuals:
   Min      1Q  Median      3Q     Max
-0.5333 -0.4043 -0.1373  0.4157  0.8487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
X            -0.3240     0.0433  -7.483 4.61e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4743 on 13 degrees of freedom
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.7971
F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06

> par(mfrow = c(2, 2))
```

> plot(concentration.lm)

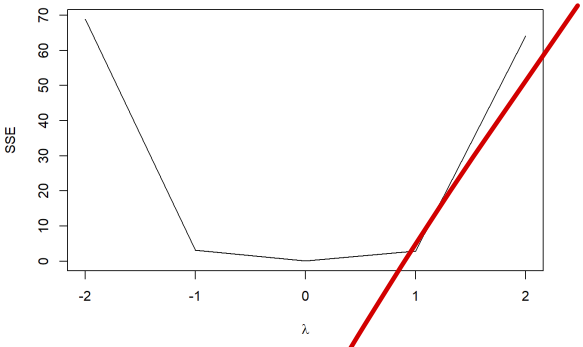


From the output, we can see that unequal error variances and nonnormality of the error terms. To remedy these departures from the simple linear regression model, we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed.

#(4)b
>library(MASS)
>library(ALSM)
>boxcox(concentration.lm,lambda = seq(-2, 2, by = 0.1))



>boxcox.sse(X, Y, l = seq(-2, 2, by = 1))



	lambda	SSE
1	-2	68.84280491
2	-1	3.16846767
5	0	0.03897303
3	1	2.92465333
4	2	64.15599383

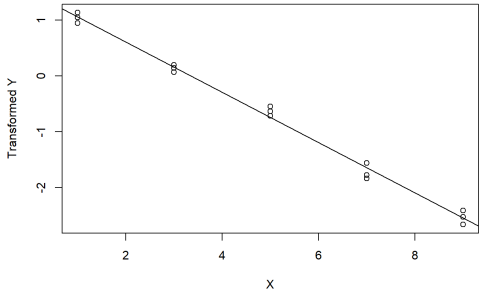
From the output, we take Y′=logeY

#(4)c
> concentration.log.lm <- lm(log(Y) ~ X, data = concentration)
> summary(concentration.log.lm)

```
Call:
lm(formula = log(Y) ~ X, data = concentration)

Residuals:
    Min      1Q  Median      3Q     Max
-0.19102 -0.10228  0.01569  0.07716  0.19699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50792    0.06028   25.01 2.22e-12 ***
X           -0.44993    0.01049  -42.88 2.19e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.115 on 13 degrees of freedom
Multiple R-squared:  0.993, Adjusted R-squared:  0.9924
F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15.
```

The least squares estimates of $\beta 0$is 1.50792 and $\beta 1$is -0.44993.
The estimated regression function is $\hat{Y}' = 1.50792 - 0.44993X$

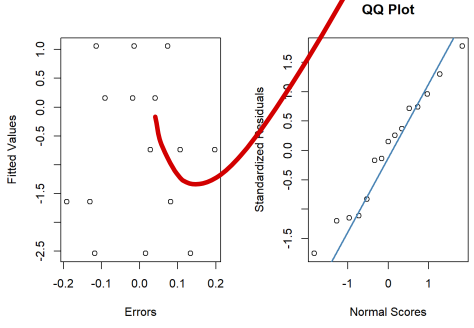
```
#(4)d
> par(mfrow = c(1, 1))
> plot(X, log(Y), xlab = "X", ylab = "Transformed Y")
> abline(concentration.log.lm)
```



From the results it was a perfect fit.

#(4)e

```
> ei <- concentration.log.lm$residuals
> yhat <- concentration.log.lm$fitted.values
> par(mfrow = c(1, 2))
> plot(ei, yhat, xlab = "Errors", ylab = "Fitted Values")
> stdei <- rstandard(concentration.log.lm)
> qqnorm(stdei, ylab = "Standardized Residuals", xlab = "Normal
Scores", main = "QQ Plot")
> qqline(stdei, col = "steelblue", lwd = 2)
```



From the output, we see that error variances are constant, errors are approximately normally distributed.

#4(f)

Since $\log_e(\hat{Y}) = 1.50792 - 0.44993X$

Hence $\hat{Y}=\exp(1.50792-0.44993X)$

4